5 个不为人知的 PDF 技巧 | 实用技巧

Minja　2018年12月31日

工作、学习中，我们都没少和 PDF 文档打交道。

PDF 能流行开来，功劳就在其**兼容性好**，本世纪的设备基本都能打开、打印时所见即所得、传给别人不用担心格式丢失或出错……以至于我和同事曾经调侃：**简单说，PDF 就是一张图片。**

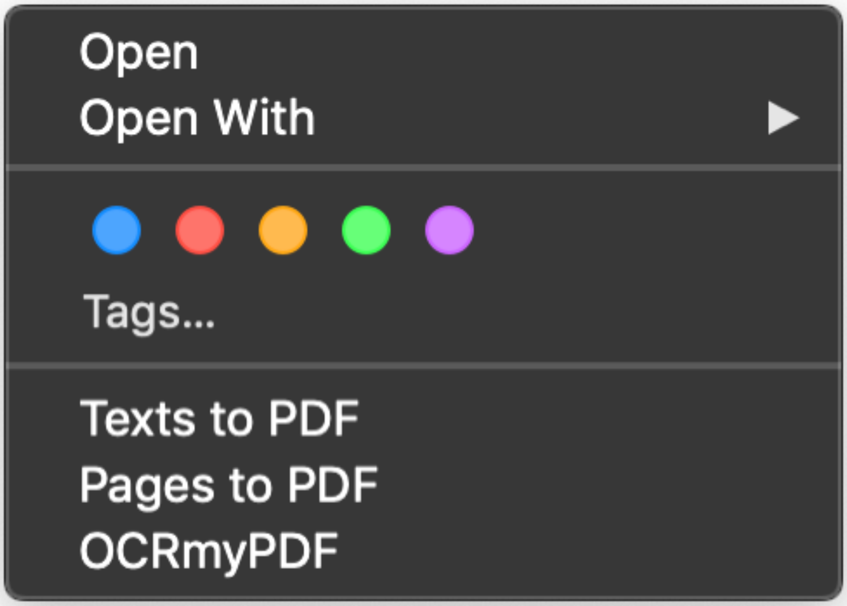
不过，如果你想对 PDF 做一些超出「图片范畴」的操作，问题就接踵而至了。这篇文章里，就收集了 5 类疑难杂症，相信不少人都遇到过：

1. **PDF 边儿太宽**，浪费空间；
2. **PDF 里全是别人的批注**，影响我阅读；
3. **PDF 居然有密码**，没法在上面批注、记笔记；
4. **扫描版 PDF 里的字没法搜索**，只能看着它干瞪眼；
5. **想把 Word 文档或其他文件转换成 PDF**，一个一个来嫌太慢。

这些问题遍布 PDF 的生成、传播到阅读、批注，而由于 PDF 的「图片」特性，我们没法像改 Word 文档一样轻松地修改 PDF，需要另辟蹊径。**本文将解决以上所有顽疾。**

上述部分问题没有现成解决方案，所以**我会提供 3 个原创的**

自动化动作。考虑到大家用的工具各不相同，我统一提供通用的 Automator 版本，安装后可以在右键菜单取用。各个第三方工具的玩家可以自行改造，做出适合自己操作习惯的动作。

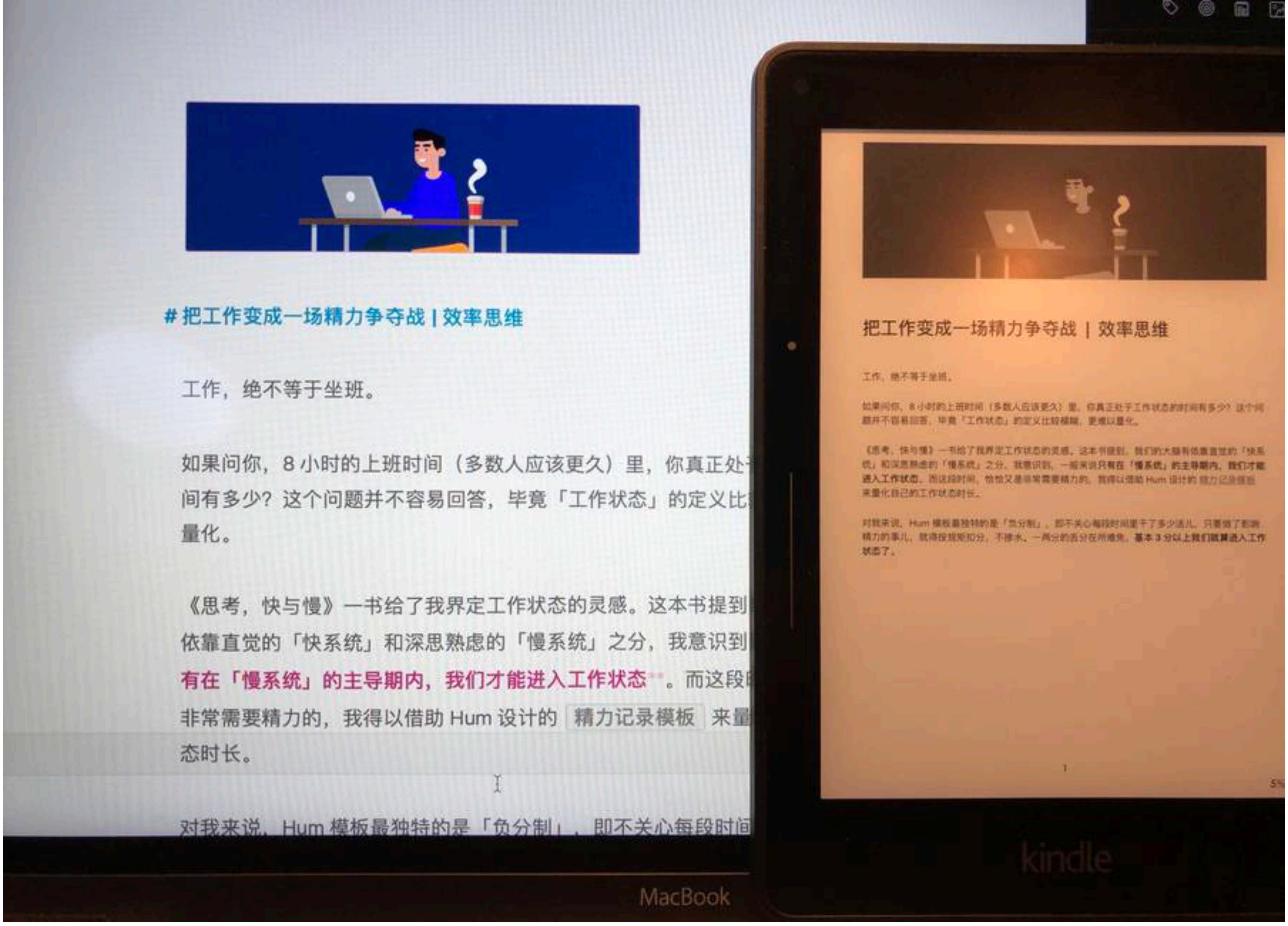

Open
Open With ▶
Tags...
Texts to PDF
Pages to PDF
OCRmyPDF

本文所有工具都可以在右键菜单取用

注：本文技巧仅供个人使用，请勿用于传播盗版资源——事实上，这些方法也入不了有批量生产需求的盗版商之眼。

技巧一：去除白边

「白边」是 PDF 文件的特色。有些白边是特意留出来放页码、页眉的，有些则是扫描时一同扫进去的白底[1]，但总而言之，它们挺很浪费屏幕空间，特别在 Kindle、小平板等小尺寸设备上，文字会被压缩好几圈。



# 把工作变成一场精力争夺战 | 效率思维

工作，绝不等于坐班。

如果问你，8 小时的上班时间（多数人应该更久）里，你真正处
间有多少？这个问题并不容易回答，毕竟「工作状态」的定义比
量化。

《思考，快与慢》一书给了我界定工作状态的灵感。这本书提到
依靠直觉的「快系统」和深思熟虑的「慢系统」之分，我意识到
有在「慢系统」的主导期内，我们才能进入工作状态**。而这段
非常需要精力的，我得以借助 Hum 设计的 精力记录模板 来量
态时长。

对我来说，Hum 模板最独特的是「负分制」，即不关心每段时间

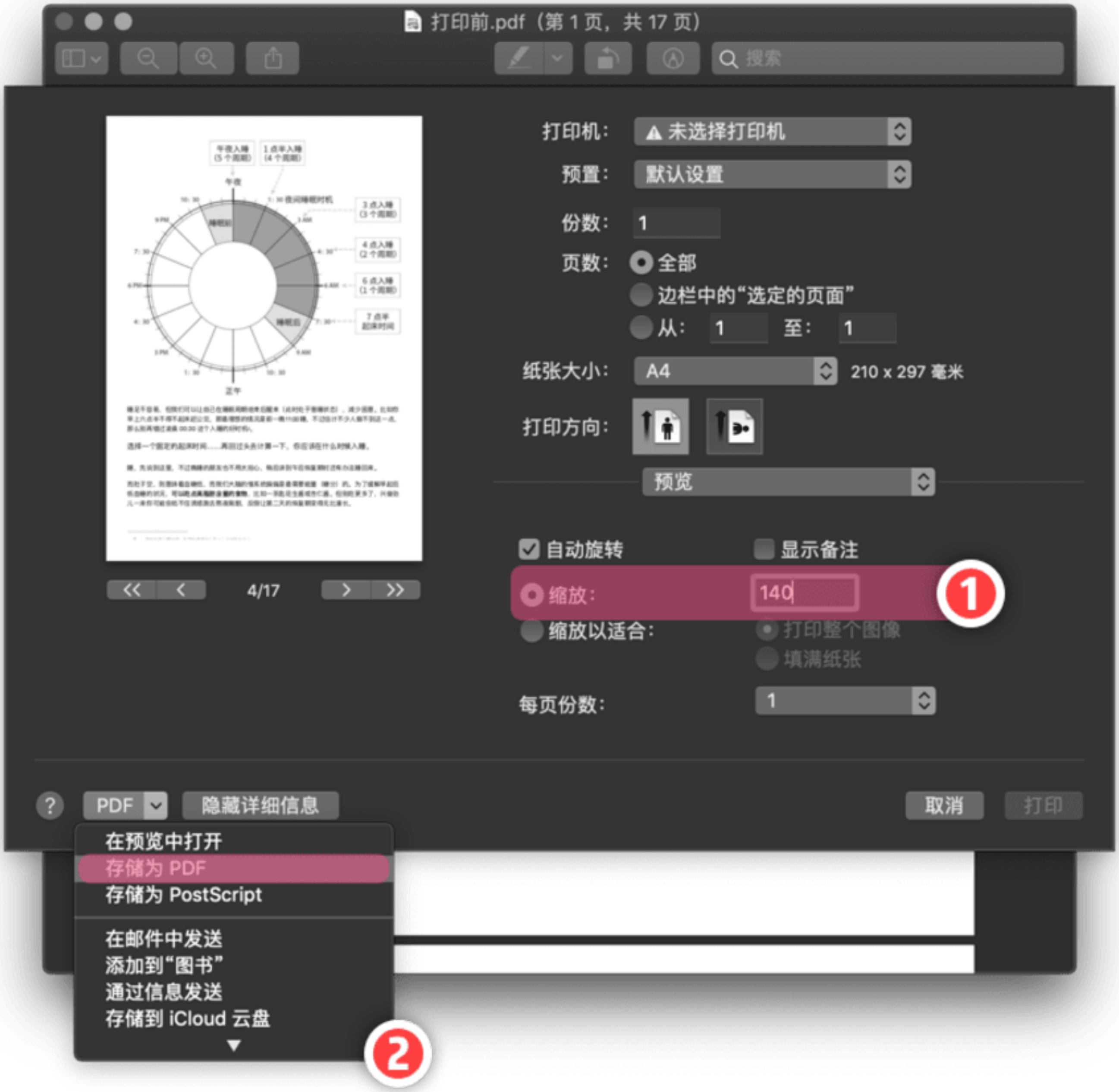
用 Kindle 看没裁过的 PDF，往往挺遭罪

去除白边有很多思路，我们用两种 macOS 原生的。**一种是重新打印，另一种是手动裁剪**，各有优劣，我们分别介绍。

重新打印

macOS 自带的打印功能可以按比例缩放 PDF，从而裁出比较整齐的新 PDF 文档。直接用预览工具打开 PDF 文件：

1. 勾选「缩放」，填入数值。留意一下左侧预览窗口，感觉白边差不多没了就行了。
2. 选择左下角「存储为 PDF」。



打印前.pdf（第 1 页，共 17 页）

打印机：未选择打印机

预置：默认设置

份数：1

页数：● 全部 ○ 边栏中的“选定的页面” ○ 从：1 至：1

纸张大小：A4 210 x 297 毫米

打印方向：

预览

☑ 自动旋转 ☐ 显示备注

● 缩放：140

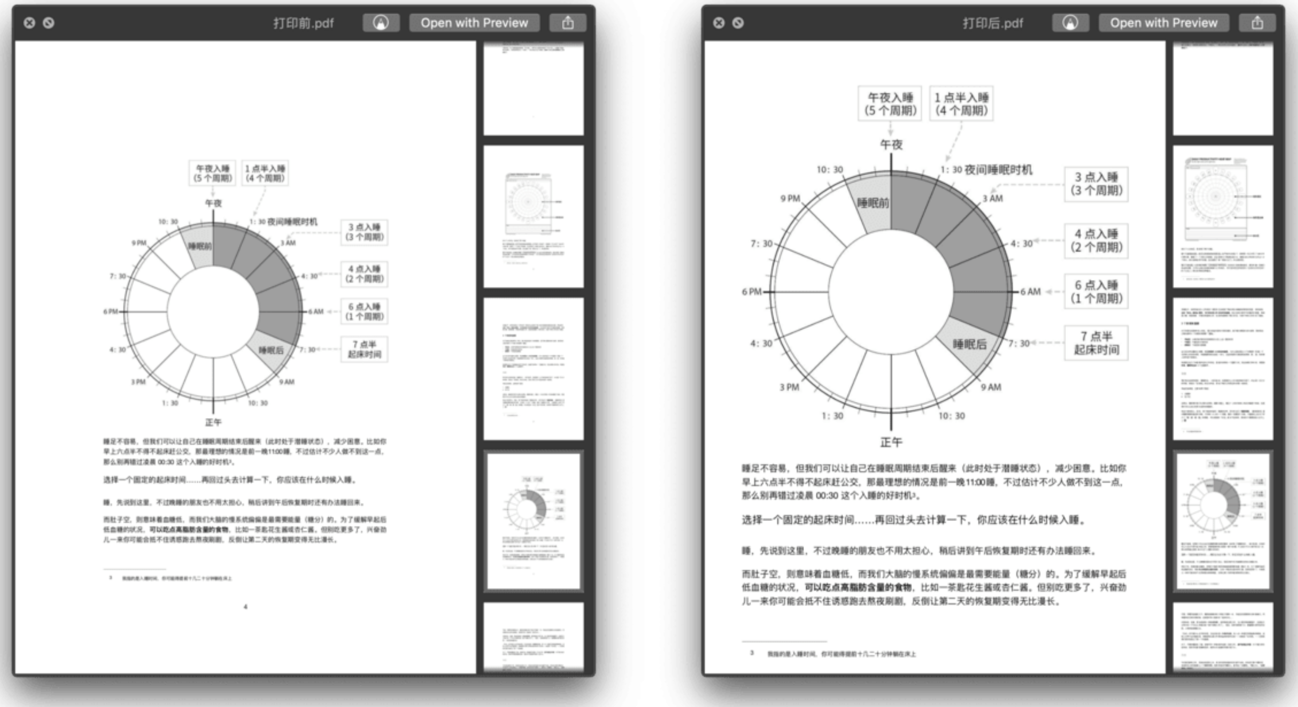
○ 缩放以适合：○ 打印整个图像 ○ 填满纸张

每页份数：1

搜索
4/17
①
?
PDF
隐藏详细信息
取消
打印
在预览中打开
存储为 PDF
存储为 PostScript
在邮件中发送
添加到“图书”
通过信息发送
存储到 iCloud 云盘
②

重新打印

重新打印后的 PDF 文件图文内容明显大了一圈，如果你需要在 Kindle 上阅读，也可以把数值填大一些，把页码一起切除。


打印前.pdf
Open with Preview
打印后.pdf
Open with Preview
午夜入睡
(5 个周期)
1 点半入睡
(4 个周期)
午夜
10: 30
1: 30 夜间睡眠时机
3 点入睡
(3 个周期)
9 PM
睡眠前
3 AM
7: 30
4: 30
4 点入睡
(2 个周期)
6 PM
6 AM
6 点入睡
(1 个周期)
4: 30
睡眠后
7: 30
7 点半
起床时间
3 PM
9 AM
1: 30
10: 30
正午

睡足不容易，但我们可以让自己在睡眠周期结束后醒来（此时处于潜睡状态），减少困意。比如你早上六点半不得不起床赶公交，那最理想的情况是前一晚11:00睡，不过估计不少人做不到这一点，那么别再错过凌晨 00:30 这个入睡的好时机[3]。

选择一个固定的起床时间……再回过头去计算一下，你应该在什么时候入睡。

睡，先说到这里，不过晚睡的朋友也不用太担心，稍后讲到午后恢复期时还有办法睡回来。

而肚子空，则意味着血糖低，而我们大脑的慢系统偏偏是最需要能量（糖分）的。为了缓解早起后低血糖的状况，**可以吃点高脂肪含量的食物**，比如一茶匙花生酱或杏仁酱。但别吃更多了，兴奋劲儿一来你可能会抵不住诱惑跑去熬夜刷剧，反倒让第二天的恢复期变得无比漫长。

[3] 我指的是入睡时间，你可能得提前十几二十分钟躺在床上

打印后白边小了很多

不过打印的缺陷也比较明显，需要 PDF 本身比较规整，可很多扫描版 PDF 的白边分布很随机，没法通过打印通吃它们[2]

，这时候就需要手动裁剪了。

手动裁剪

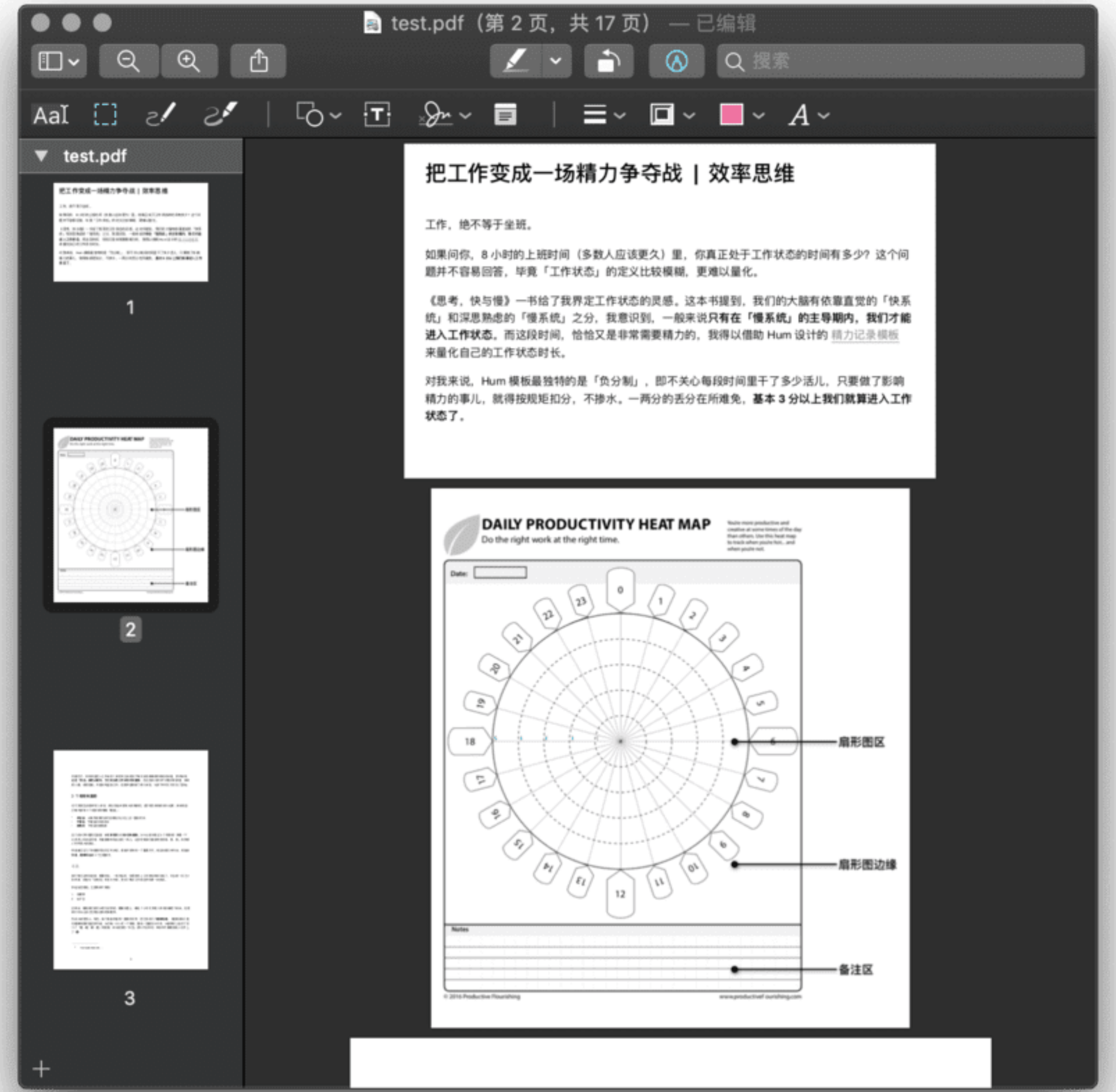
听到「手动」两字不用怕，**其实还是可以批量处理的**。仍然用预览工具打开 PDF：

1. 点击左上角的按钮，开启「缩略图」视图；
2. 按下 `⌘Command - A`，选中所有页面；
3. 点击「菜单栏 - 工具 - 矩形选择」；
4. 选中需要留下的区域，按 `⌘Command - K` 进行裁剪。

裁剪前可以发现，尽管只在 PDF 其中一页进行了选取，但选

区被应用到了所有页面——因为我们已经提取选中了它们，**从而实现批量处理**。开一方面，相比重新打印，裁剪还可以更灵活：

1. 选区能自行调整，不一定非得是 A4 纸[3] 的长宽比；
2. 如果部分页面如果裁不齐，可以单独再行修剪。


test.pdf（第 2 页，共 17 页）— 已编辑
搜索
AaI
test.pdf
1
2
3
DAILY PRODUCTIVITY HEAT MAP
Do the right work at the right time.
Date:
Notes
扇形图区
扇形图边缘
备注区

把工作变成一场精力争夺战 | 效率思维

工作，绝不等于坐班。

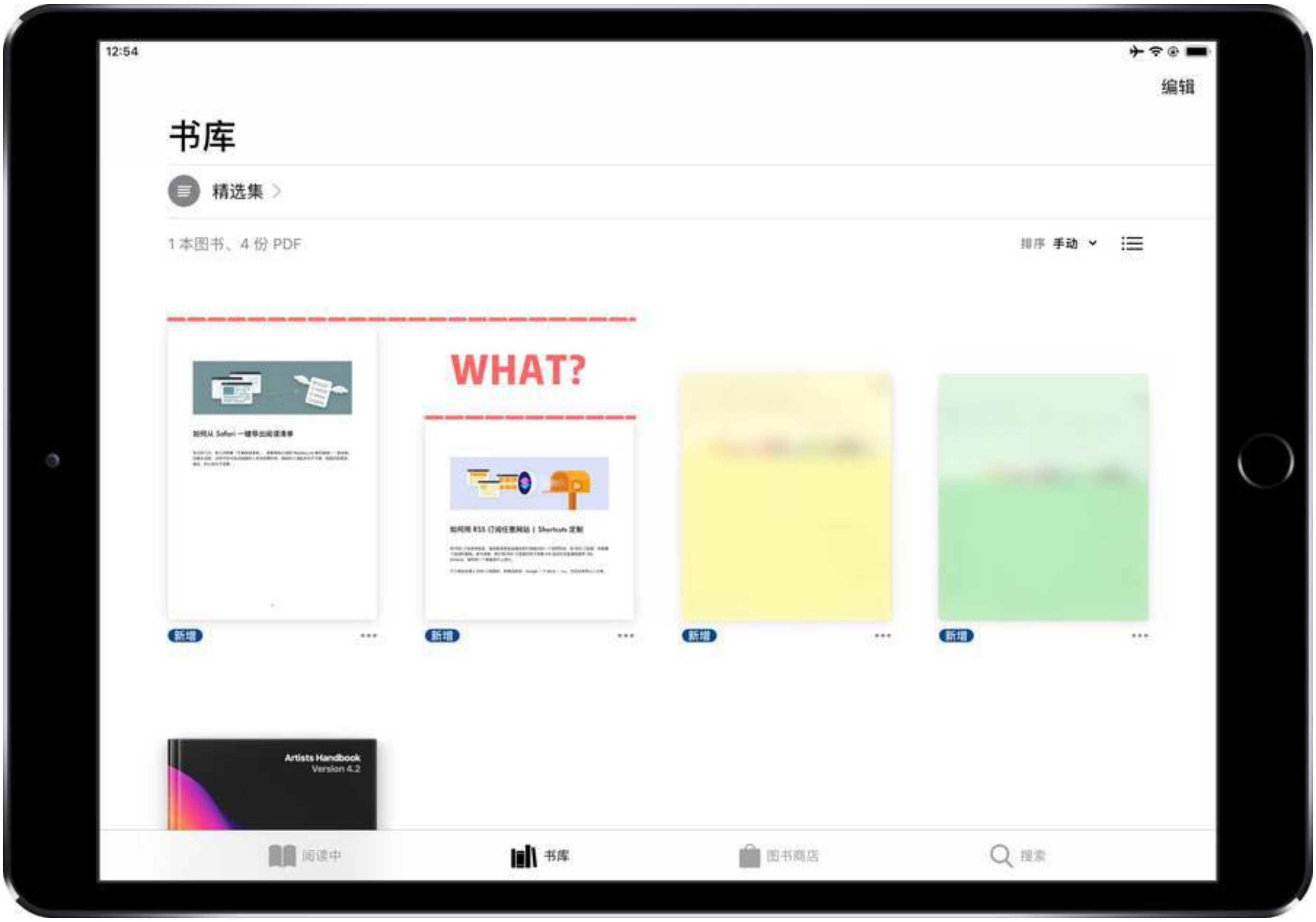
如果问你，8 小时的上班时间（多数人应该更久）里，你真正处于工作状态的时间有多少？这个问题并不容易回答，毕竟「工作状态」的定义比较模糊，更难以量化。

《思考，快与慢》一书给了我界定工作状态的灵感。这本书提到，我们的大脑有依靠直觉的「快系统」和深思熟虑的「慢系统」之分，我意识到，一般来说**只有在「慢系统」的主导期内，我们才能进入工作状态**。而这段时间，恰恰又是非常需要精力的。我得以借助 Hum 设计的 精力记录模板 来量化自己的工作状态时长。

对我来说，Hum 模板最独特的是「负分制」，即不关心每段时间里干了多少活儿，只要做了影响精力的事儿，就得按规矩扣分，不掺水。一两分的丢分在所难免，**基本 3 分以上我们就算进入工作状态了。**

手动裁剪时，不同页面可以裁成不同大小

不过手动裁剪也有问题，**首先就是很难裁整齐。**如果你的阅读器能够显示 PDF 封面（基本都能显示，逃不掉的），那你将获得一排参差不齐的书籍封面，看着很碍眼。


12:54
编辑
书库
精选集
1 本图书、4 份 PDF
排序 手动
WHAT?
新增
新增
新增
新增
Artists Handbook
Version 4.2
阅读中
书库
图书商店
搜索

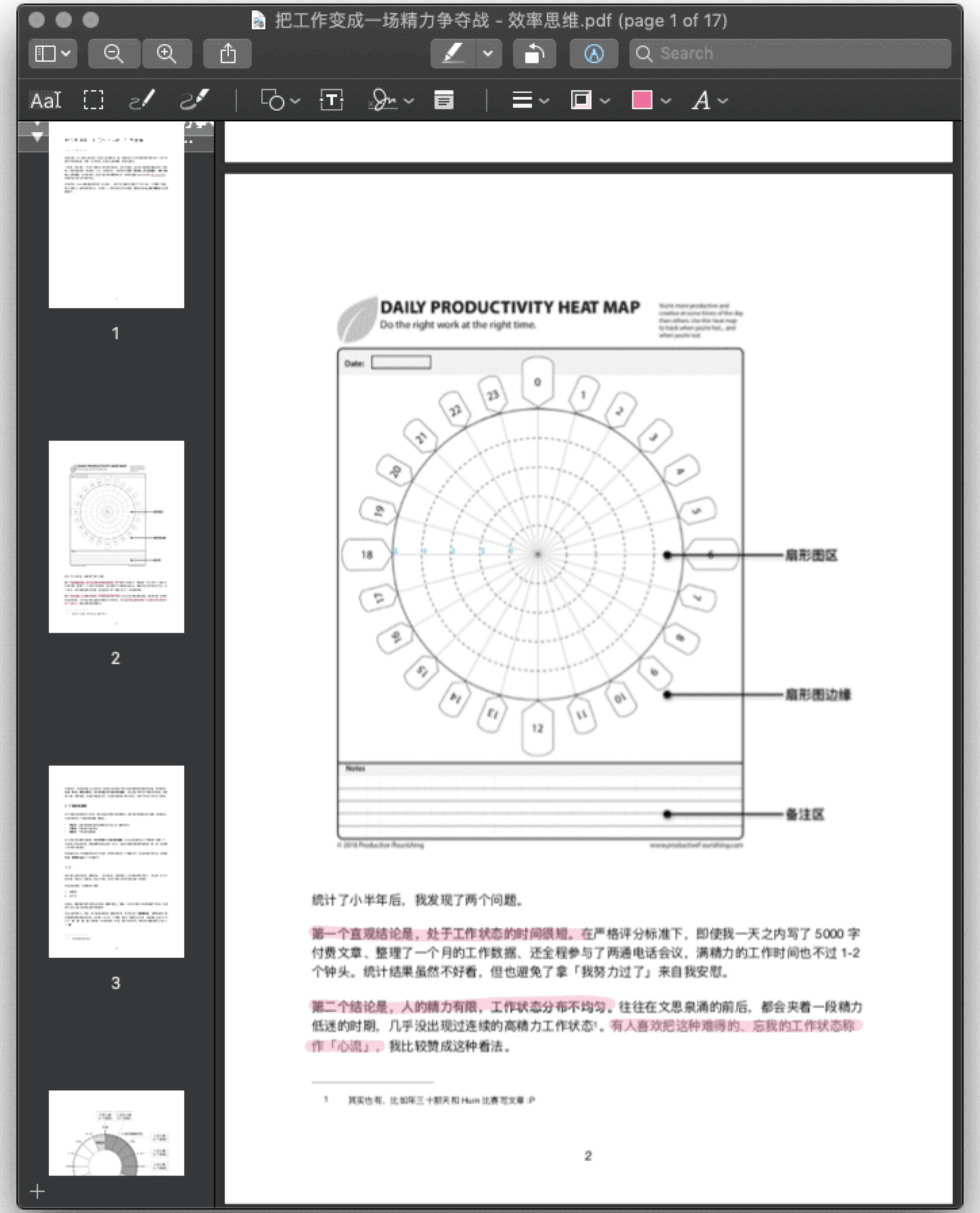
高低不齐的封面看着很难受

另外，扫描版的 PDF 文件往往体积巨大，动几十上百兆，而裁剪操作仅是隐藏边上的部分，并没有删除它们，**裁好后文件体积不会减小。**

总之，如果想裁得整齐、精巧，又不在意目录，那可以重新打印；如果需要保留目录，那还是手动裁剪。

技巧二：去除批注

有些 PDF 流传了几代，交到我们手里时已经画满了前人的标记。这些用功的痕迹，很大程度上也影响了正常的阅读、批注。


把工作变成一场精力争夺战 - 效率思维.pdf (page 1 of 17)
Search
AaI
1
2
3
DAILY PRODUCTIVITY HEAT MAP
Do the right work at the right time.
Date:
扇形图区
扇形图边缘
Notes
备注区

统计了小半年后，我发现了两个问题。

第一个直观结论是，处于工作状态的时间很短。在严格评分标准下，即使我一天之内写了 5000 字付费文章、整理了一个月的工作数据、还全程参与了两通电话会议，满精力的工作时间也不过 1-2 个钟头。统计结果虽然不好看，但也避免了拿「我努力过了」来自我安慰。

第二个结论是，人的精力有限，工作状态分布不均匀。往往在文思泉涌的前后，都会夹着一段精力低迷的时期，几乎没出现过连续的高精力工作状态[1]。有人喜欢把这种难得的、忘我的工作状态称作「心流」，我比较赞成这种看法。

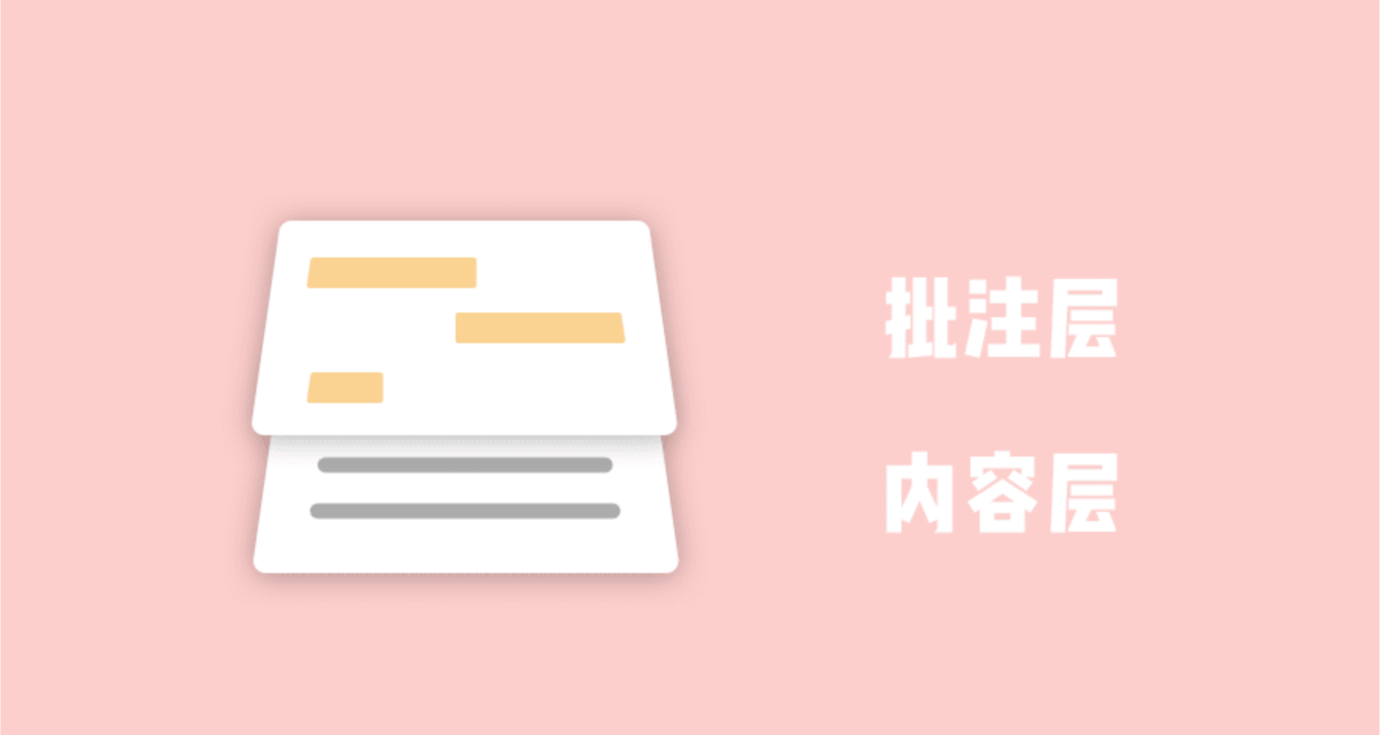
---

[1] 其实也有，比如年三十那天和 Hum 比赛写文章 :P

2

画满了批注的PDF

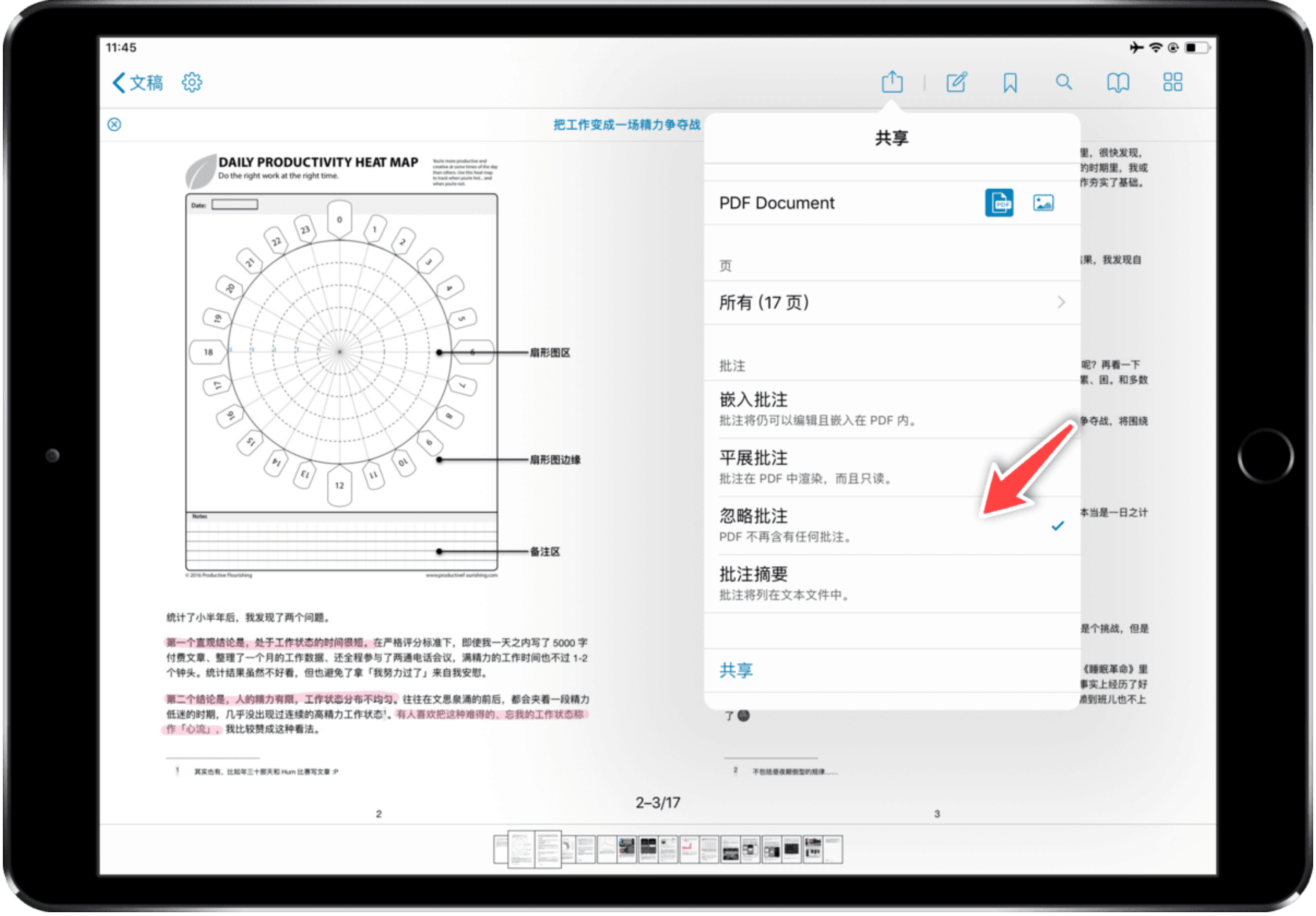
虽说 PDF 可以看作一张「图」，但它可能也是分层的，批注就是最顶端的「图层」，能够选中后删除。



PDF 其实是分层的

不过文档里的批注出没无常，就算有耐性一个一个手动删，也不一定拔得干净。PDF Expert、Adobe PDF Acrobat 这些专业工具倒是能够批量消灭批注，但它们价格实在太高，这里介绍一个不花钱的工具：**iOS 上的 PDF Viewer**。

PDF Viewer 是那种讨人喜欢的小巧阅读器，原生界面，简单的导入导出功能、必要的批注工具，最重要的是可以导出纯净版 PDF，不含标注。



11:45
文稿
把工作变成一场精力争夺战
DAILY PRODUCTIVITY HEAT MAP
Do the right work at the right time.
扇形图区
扇形图边缘
备注区
2–3/17
2
3
共享
PDF Document
页
所有 (17 页)
批注
嵌入批注
批注将仍可以编辑且嵌入在 PDF 内。
平展批注
批注在 PDF 中渲染，而且只读。
忽略批注
PDF 不再含有任何批注。
批注摘要
批注将列在文本文件中。
共享

统计了小半年后，我发现了两个问题。

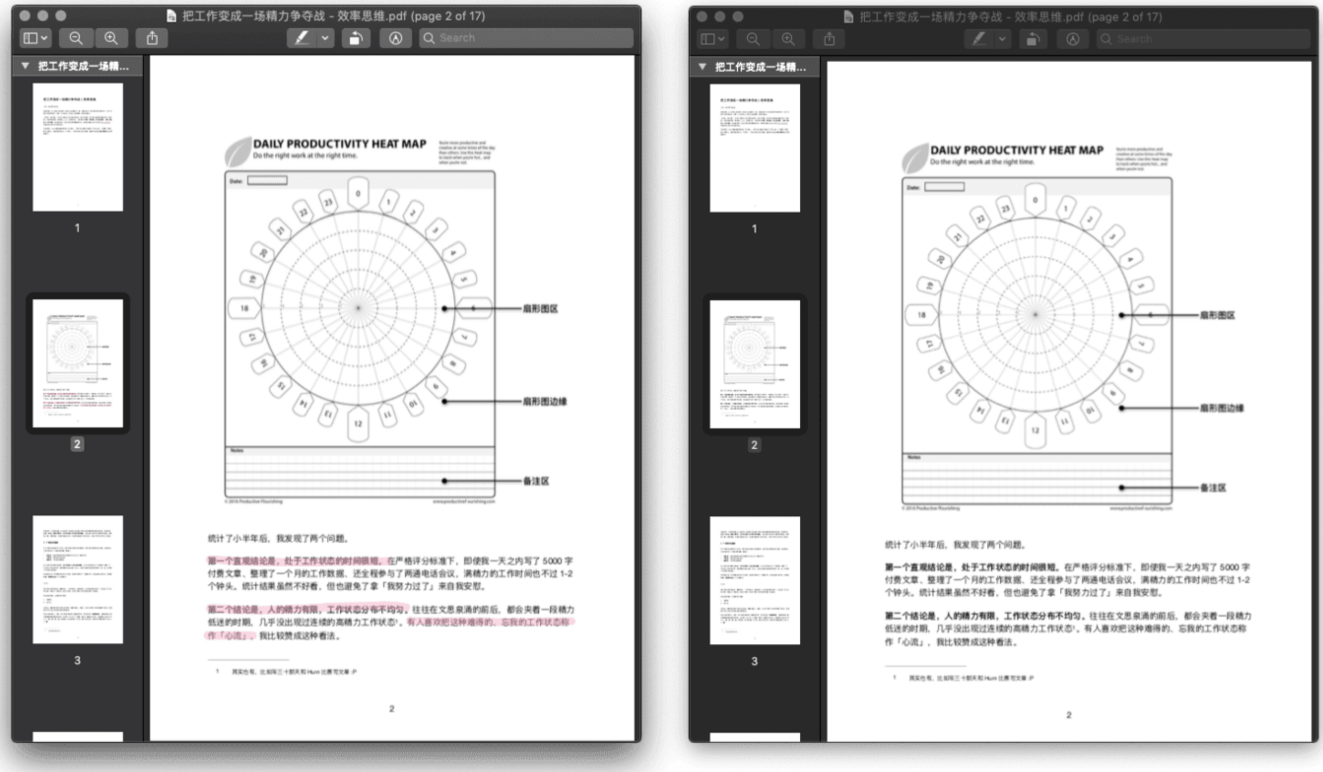
第一个直观结论是，处于工作状态的时间很短。在严格评分标准下，即使我一天之内写了 5000 字付费文章、整理了一个月的工作数据、还全程参与了两通电话会议，满精力的工作时间也不过 1-2 个钟头。统计结果虽然不好看，但也避免了拿「我努力过了」来自我安慰。

第二个结论是，人的精力有限，工作状态分布不均匀。往往在文思泉涌的前后，都会夹着一段精力低迷的时期，几乎没出现过连续的高精力工作状态[1]。有人喜欢把这种难得的、忘我的工作状态称作「心流」，我比较赞成这种看法。

[1] 其实也有，比如年三十那天和 Hum 比赛写文章 :P

PDF Viewer 可以导出不带批注的 PDF

没有多余的步骤，一份干净的 PDF 就出现了。


把工作变成一场精力争夺战 - 效率思维.pdf (page 2 of 17)
Search
DAILY PRODUCTIVITY HEAT MAP
Do the right work at the right time.
扇形图区
扇形图边缘
备注区
Notes

统计了小半年后，我发现了两个问题。

第一个直观结论是，处于工作状态的时间很短。在严格评分标准下，即使我一天之内写了 5000 字付费文章、整理了一个月的工作数据、还全程参与了两通电话会议，满精力的工作时间也不过 1-2 个钟头。统计结果虽然不好看，但也避免了拿「我努力过了」来自我安慰。

第二个结论是，人的精力有限，工作状态分布不均匀。往往在文思泉涌的前后，都会夹着一段精力低迷的时期，几乎没出现过连续的高精力工作状态[1]。有人喜欢把这种难得的、忘我的工作状态称作「心流」，我比较赞成这种看法。

[1] 其实也有，比如年三十那天和 Hum 比赛写文章 :P

2

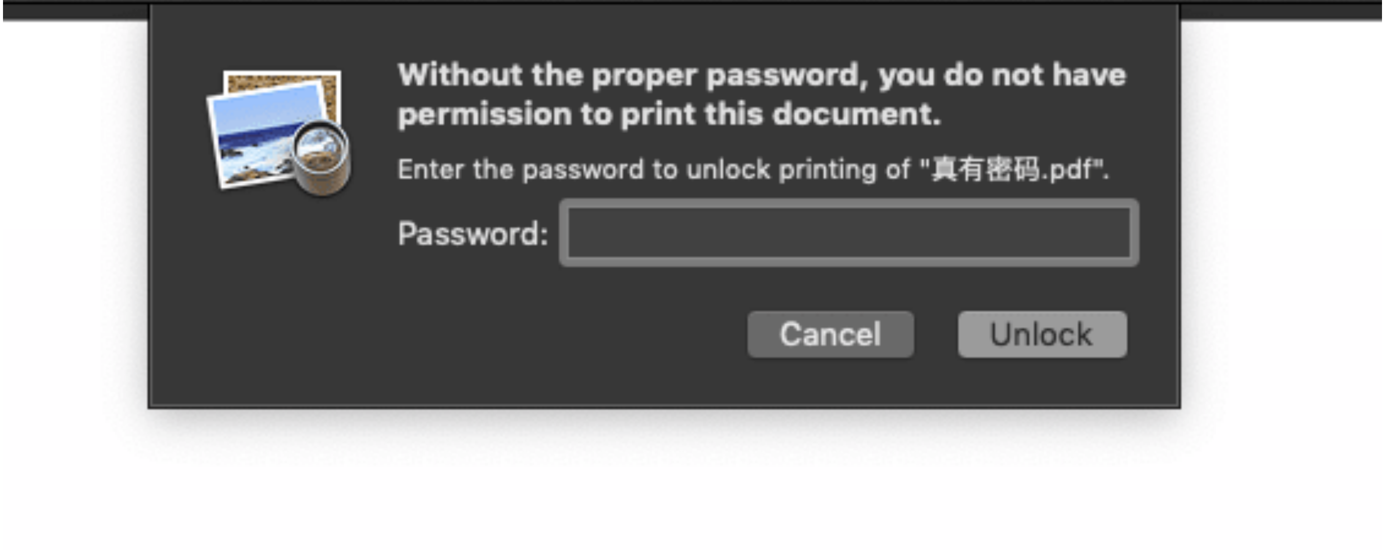
去除批注前后

对了，如果你下载了一份插满推广链接的 PDF——尤其是那种不小心碰一下就会弹到浏览器的烦人链接——上述方法或许也能起到批量清理作用。

技巧三：绕过编辑密码

有些 PDF 能打开阅读，但是不能编辑，做笔记也不行，很头疼。这就是遇到拥有者密码（Owner Permission Password）了。

初衷是好的，但有时也会给自己人带来不便。读书时就有这种情况，系里祖传的一些 PDF 书籍，当年的制作者不知何处高就去，只留密码挡我路，在我们看到精彩处、大笔一挥想记点东西时，发现没有权限编辑。



Without the proper password, you do not have permission to print this document.

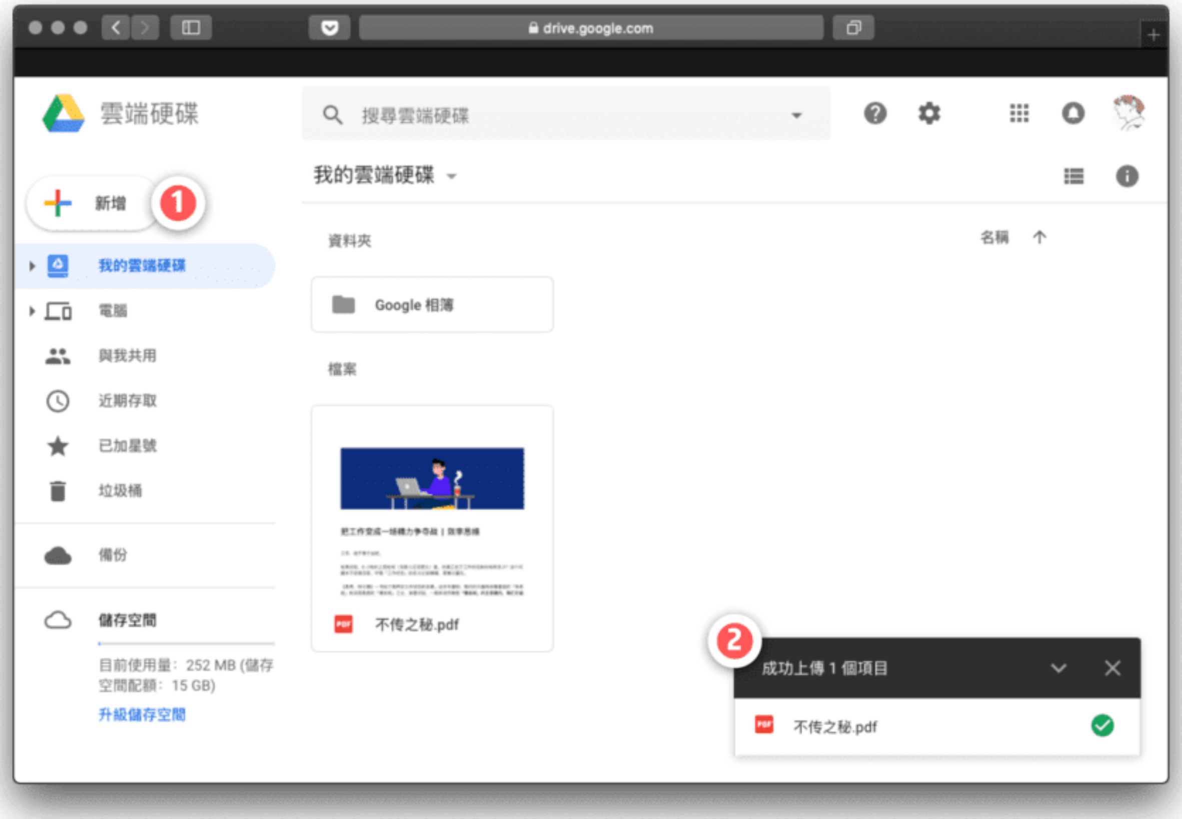
Enter the password to unlock printing of "真有密码.pdf".

Password:

编辑密码严重影响做笔记

好在这类编辑密码也有办法绕过的：不用什么「破解」，**只需要一个 Google Drive 网盘**。

1. 先确认一遍，你遇到的的确是拥有者密码，如果连打开都不给打开，你还是找提供 PDF 的人讨密码吧；确认后，**把想编辑的 PDF 文档上传到 Google Drive**。到这一步，事儿就成了一半了。


drive.google.com
雲端硬碟
搜尋雲端硬碟
新增
1
我的雲端硬碟
電腦
與我共用
近期存取
已加星號
垃圾桶
備份
儲存空間
升級儲存空間
我的雲端硬碟
資料夾
名稱
Google 相簿
檔案
2

目前使用量：252 MB (儲存空間配額：15 GB)

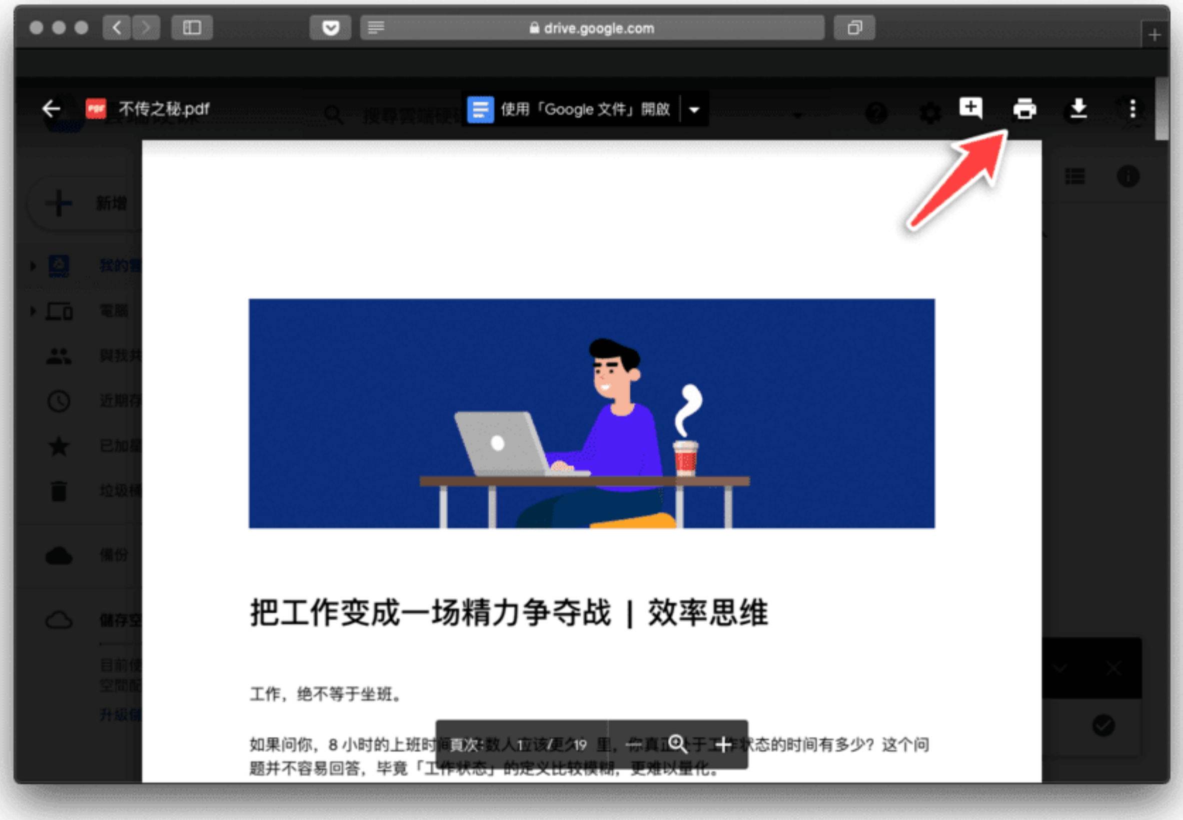
不传之秘.pdf

成功上傳 1 個項目

不传之秘.pdf

上传 PDF 到 Google Drive

2. 接着就省力了，在 Google Drive 里点击查看 PDF，**点击右上角的打印按钮**，静待一分钟，不带密码的 PDF 就生成了，Google Drive 会帮你打开一个预览的页面。


drive.google.com
不传之秘.pdf
使用「Google 文件」開啟
新增
頁次 1 / 19

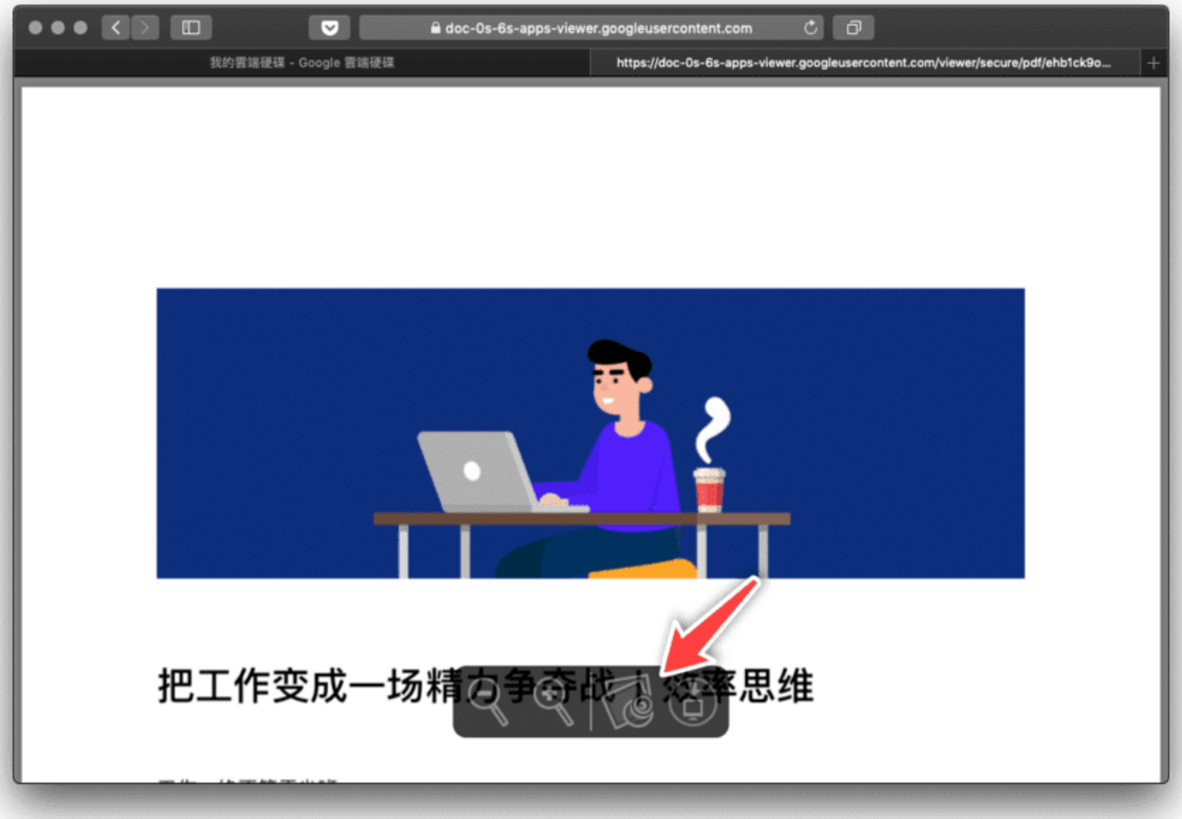
把工作变成一场精力争夺战 | 效率思维

工作，绝不等于坐班。

如果问你，8 小时的上班时[illegible]里，[illegible]态的时间有多少？这个问题并不容易回答，毕竟「工作状态」的定义比较模糊，更难以量化。

在 Google Drive 中打印 PDF

3. **最后点击预览界面的下载按钮**，新 PDF 就躺在了你的下载文件夹。如果没看到下载按钮，就在页面中下方晃几下鼠标把它召唤出来。


doc-0s-6s-apps-viewer.googleusercontent.com
我的雲端硬碟 - Google 雲端硬碟
https://doc-0s-6s-apps-viewer.googleusercontent.com/viewer/secure/pdf/ehb1ck9o...

把工作变成一场精力争夺战！效率思维

下载没有密码的 PDF 到本地

说到底，这个技巧的原理只是把你看到动东西重新打印了一遍，其他网盘或在线文档协作工具可能也支持，大家不妨自己测试一下，如果有能在国内可用的服务就更好了。

之前坊间流传过用 Chrome 浏览器重新打印来「解锁」密码的方子，不过在最新版本的浏览器中已经失效。我做了些测试，当前版本的 Chromium、Firefox、Safari 等浏览器均无法直接打印带密码的 PDF。

此外，网上还有一些破解 PDF 密码的站点，需要上传原始 PDF 文件，无论出于版权还是安全因素考虑，这里并不推荐[4]

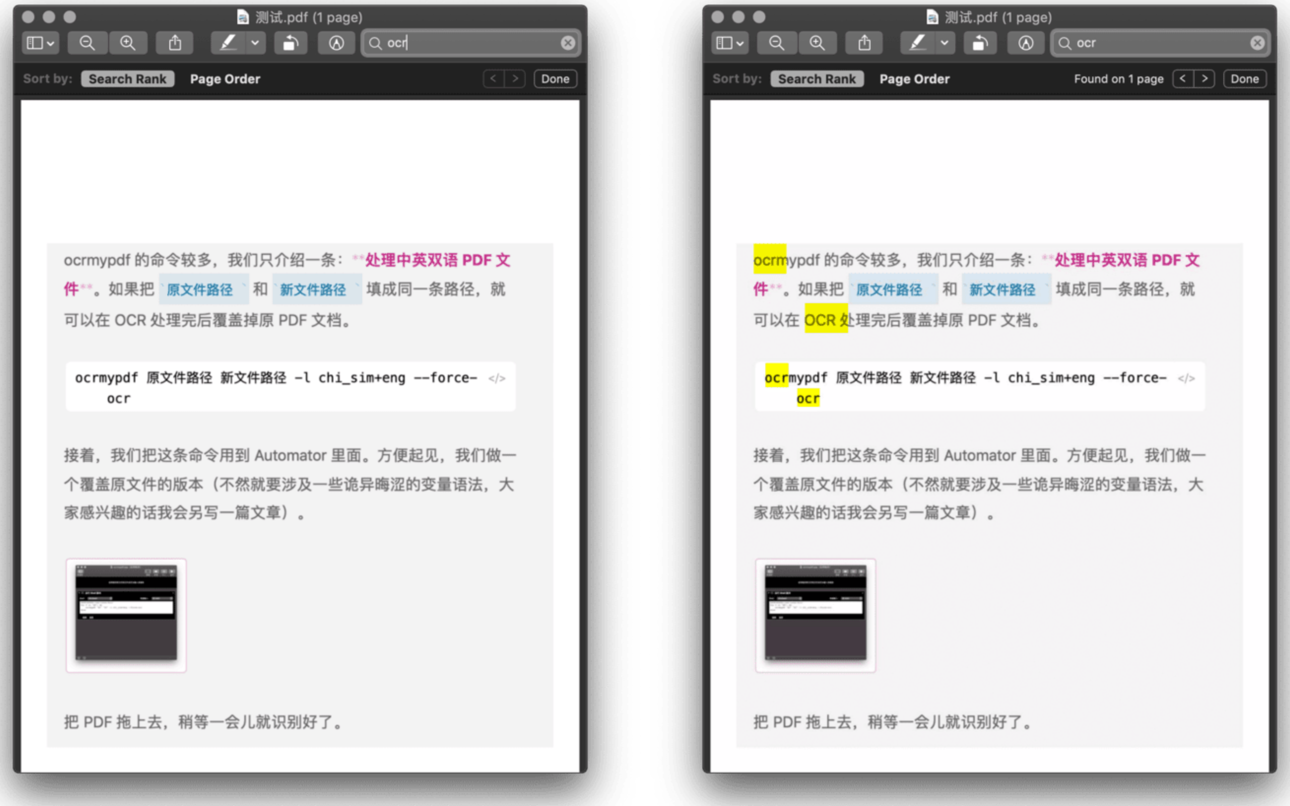
。当然，如果是一些从不可言说网站下载来的 PDF，连解

压/打开密码都不知道，那就无能为力了 🤷‍♂️

技巧四：将扫描版 PDF 转换成可搜索的文档

读过扫描版 PDF 的人，都知道那种**能看不能搜**的感觉是多么难受。不巧，有些老教材又只有扫描版，似乎只好硬着头皮人肉检索。

@契丹神童 曾经在 Power+ 1.0 介绍过 ocrmypdf 这个命令行工具，**能够给扫描件填上一层文字层，使得文本能够被搜索**；我据此做了一个 Automator 动作，这样，不用买第三方工具也可以享受 OCR 功能。



ocrmypdf 的命令较多，我们只介绍一条：**处理中英双语 PDF 文件**。如果把 `原文件路径` 和 `新文件路径` 填成同一条路径，就可以在 OCR 处理完后覆盖掉原 PDF 文档。

```
ocrmypdf 原文件路径 新文件路径 -l chi_sim+eng --force-
    ocr
```

接着，我们把这条命令用到 Automator 里面。方便起见，我们做一个覆盖原文件的版本（不然就要涉及一些诡异晦涩的变量语法，大家感兴趣的话我会另写一篇文章）。

把 PDF 拖上去，稍等一会儿就识别好了。

ocrmypdf 的命令较多，我们只介绍一条：**处理中英双语 PDF 文件**。如果把 `原文件路径` 和 `新文件路径` 填成同一条路径，就可以在 OCR 处理完后覆盖掉原 PDF 文档。

ocrmypdf 原文件路径 新文件路径 -l chi_sim+eng --force-ocr

接着，我们把这条命令用到 Automator 里面。方便起见，我们做一个覆盖原文件的版本（不然就要涉及一些诡异晦涩的变量语法，大家感兴趣的话我会另写一篇文章）。

把 PDF 拖上去，稍等一会儿就识别好了。

ocrmypdf 处理过的文件能被搜索

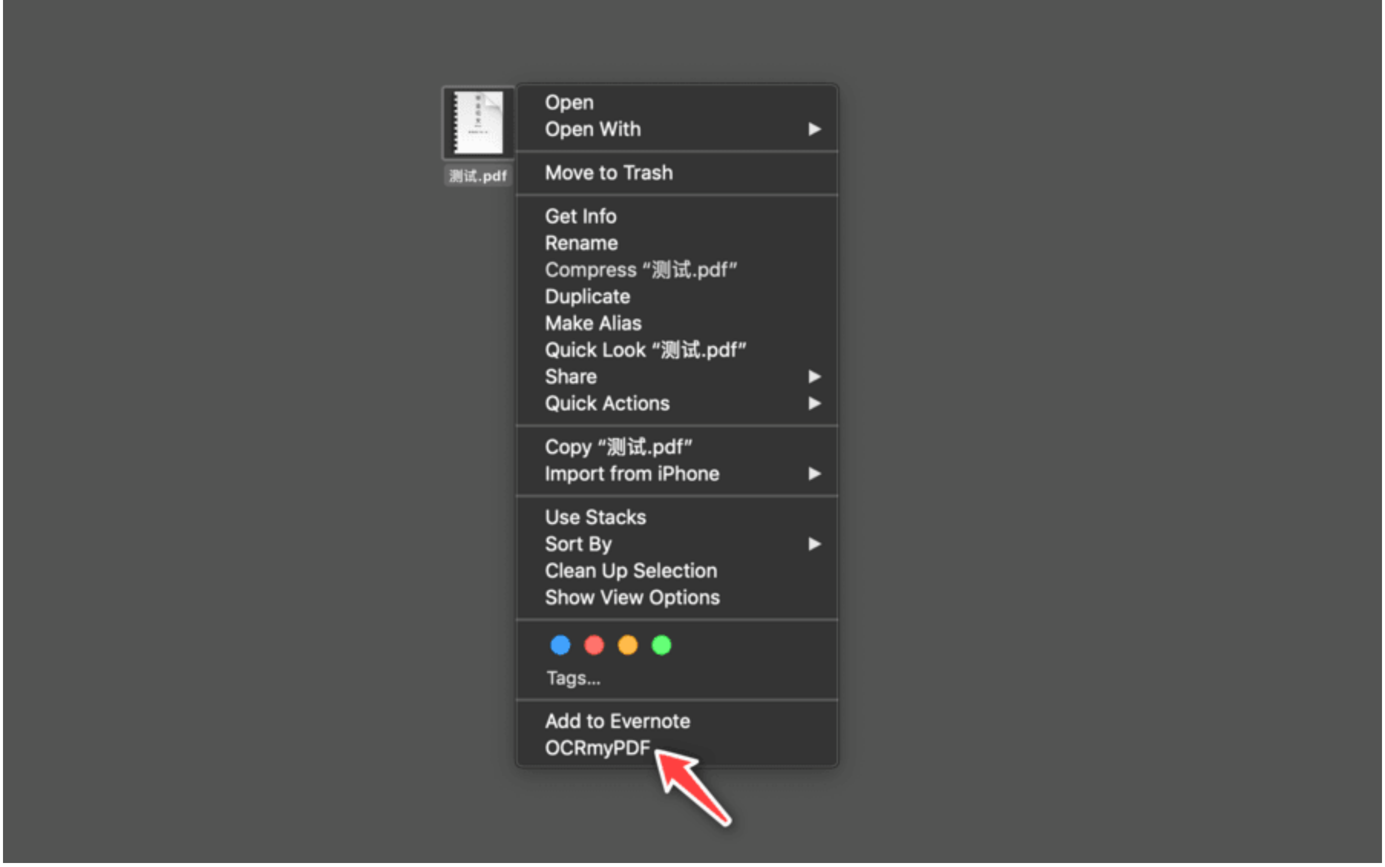
> 动作下载 🔗

光装好动作，还不能直接用，需要安装和配置一下 ocrmypdf。**下面是安装 ocrmypdf 及中文语言包的脚本，直接复制到终端中运行即可。**截至写这篇文章，tesseract 的最新版本是 3.05.02，如果下载语言包失败，请到 `/usr/local/Cellar/tesseract/` 下检查一下你的版本是多少，并把命令里的旧版本号 `3.05.02` 替换掉。

```
# 安装脚本
brew install ocrmypdf
wget -o
/usr/local/Cellar/tesseract/3.05.02/share/tessdata/chi_sim.tr
 https://github.com/tesseract-
ocr/tessdata/raw/master/chi_sim.traineddata
```

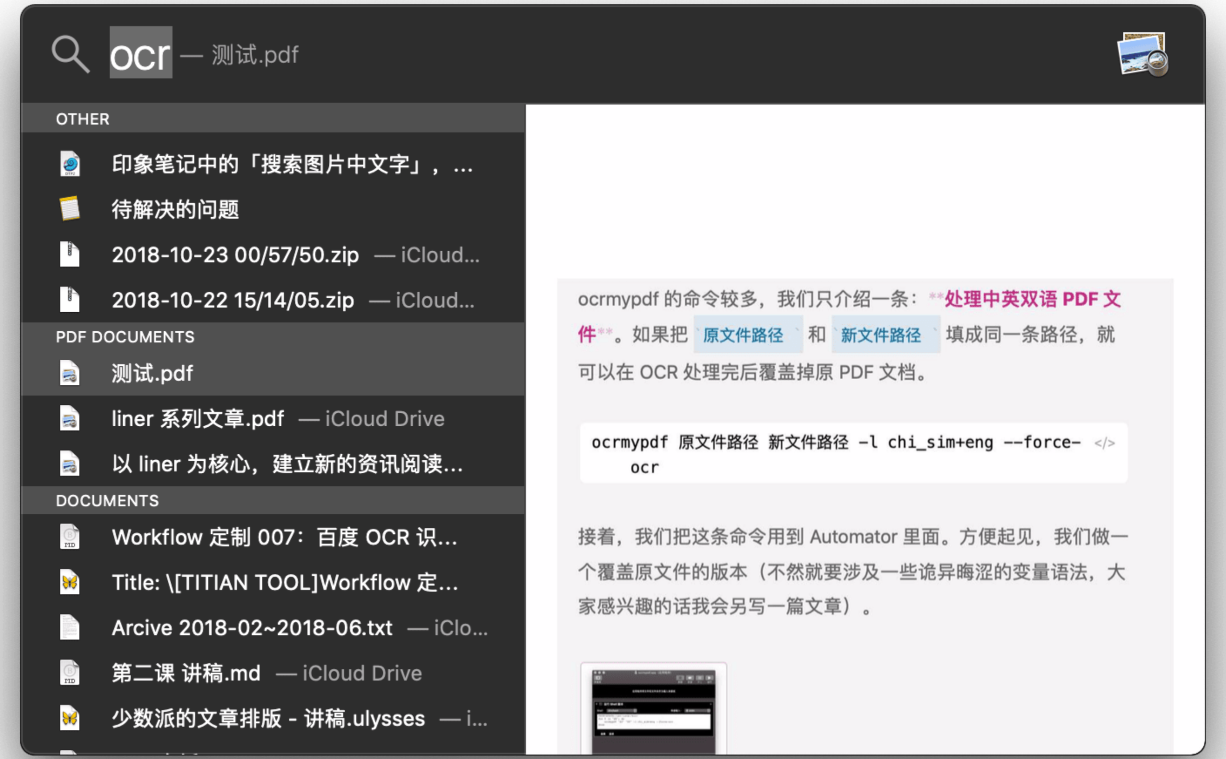
装好动作、配置完成后，OCR 功能就会出现在右键菜单里。

挑一个扫描版或图片版本的 PDF 识别一下，短则几秒钟，长则几分钟，具体要看 PDF 大小。


测试.pdf
Open
Open With
Move to Trash
Get Info
Rename
Compress "测试.pdf"
Duplicate
Make Alias
Quick Look "测试.pdf"
Share
Quick Actions
Copy "测试.pdf"
Import from iPhone
Use Stacks
Sort By
Clean Up Selection
Show View Options
Tags...
Add to Evernote
OCRmyPDF

OCR 功能在右键菜单里

动作运行结束后再打开 PDF，就能进行搜索了。其实不光在 PDF 文档里面可以搜索，直接用 Spotlight、Finder 也能搜刚才处理过的 PDF，很方便。



ocr — 测试.pdf

OTHER
印象笔记中的「搜索图片中文字」，...
待解决的问题
2018-10-23 00/57/50.zip — iCloud...
2018-10-22 15/14/05.zip — iCloud...
PDF DOCUMENTS
测试.pdf
liner 系列文章.pdf — iCloud Drive
以 liner 为核心，建立新的资讯阅读...
DOCUMENTS
Workflow 定制 007：百度 OCR 识...
Title: \[TITIAN TOOL]Workflow 定...
Arcive 2018-02~2018-06.txt — iClo...
第二课 讲稿.md — iCloud Drive
少数派的文章排版 - 讲稿.ulysses — i...

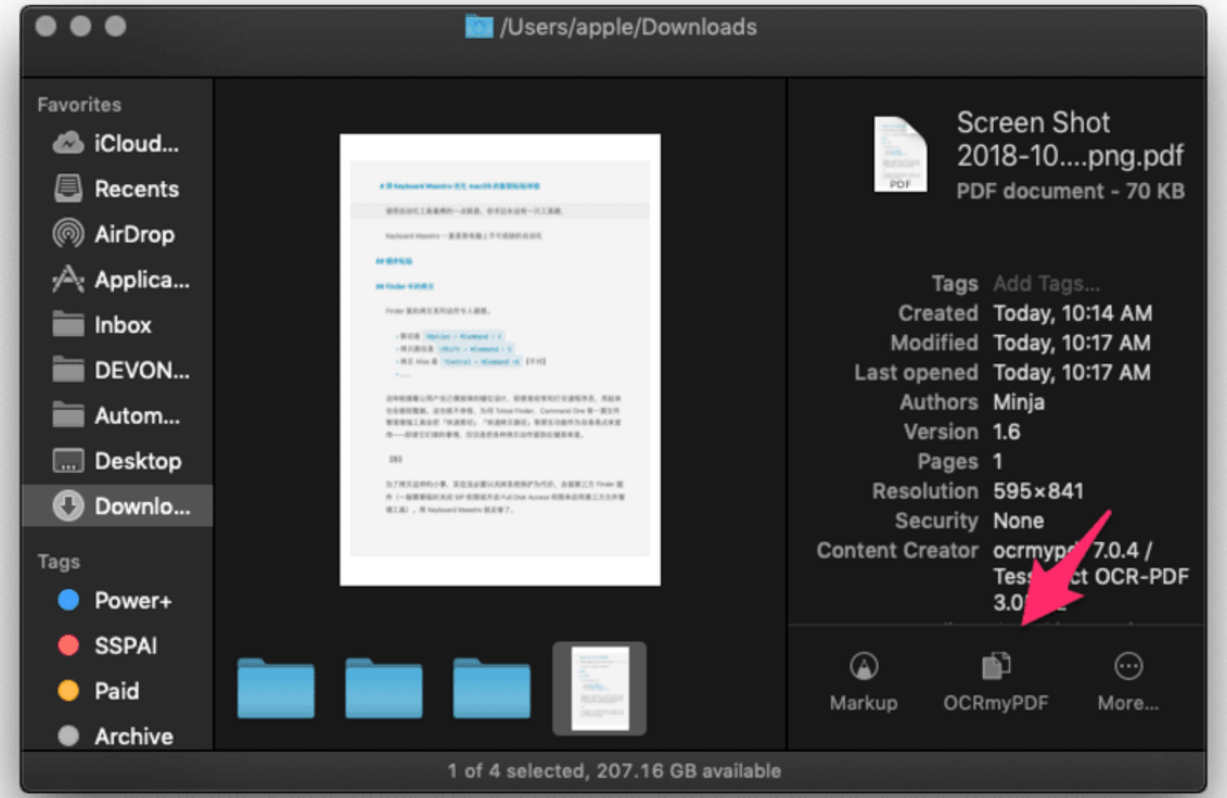
ocrmypdf 的命令较多，我们只介绍一条：**处理中英双语 PDF 文件**。如果把 `原文件路径` 和 `新文件路径` 填成同一条路径，就可以在 OCR 处理完后覆盖掉原 PDF 文档。

```
ocrmypdf 原文件路径 新文件路径 -l chi_sim+eng --force-
    ocr
```

接着，我们把这条命令用到 Automator 里面。方便起见，我们做一个覆盖原文件的版本（不然就要涉及一些诡异晦涩的变量语法，大家感兴趣的话我会另写一篇文章）。

通过 Spotlight 搜索到了 ocrmypdf 处理过的文件

除了右键菜单，刚才你下载的 Automator Workflow 还可以当作快速操作来用，直接在 Finder 的 PDF 预览面板里进行 OCR 识别。


/Users/apple/Downloads
Favorites
iCloud...
Recents
AirDrop
Applica...
Inbox
DEVON...
Autom...
Desktop
Downlo...
Tags
Power+
SSPAI
Paid
Archive
Markup
OCRmyPDF
More...

Screen Shot 2018-10....png.pdf

PDF document - 70 KB

Tags	Add Tags...
Created	Today, 10:14 AM
Modified	Today, 10:17 AM
Last opened	Today, 10:17 AM
Authors	Minja
Version	1.6
Pages	1
Resolution	595×841
Security	None
Content Creator	ocrmyp 7.0.4 / Tess ct OCR-PDF 3.0

1 of 4 selected, 207.16 GB available

把 OCR 动作集成到 Finder

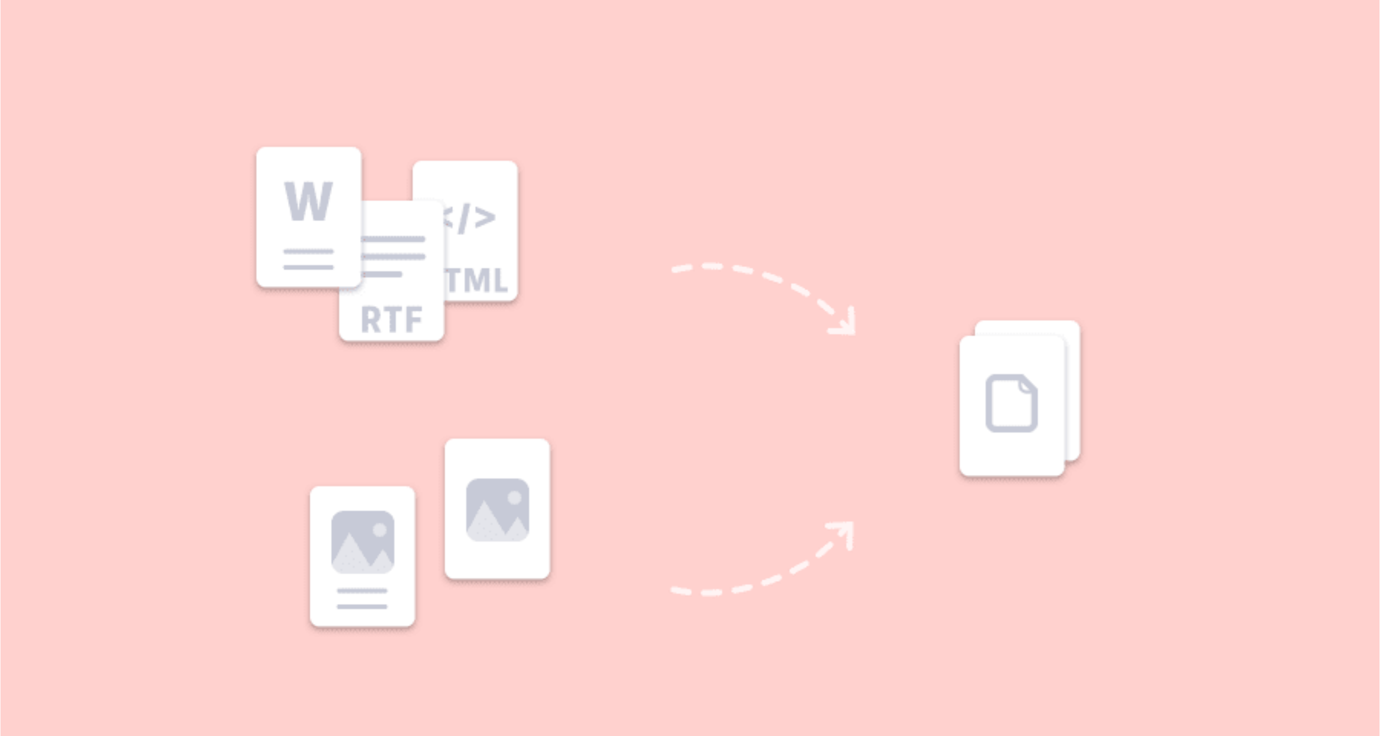
ocrmypdf 的功能非常强大，我只用到了其中的中文和英文 OCR 功能，如果你还要和其他语种打交道的话，不妨读读契丹神童的文章，试着自己配置语言包。

推荐阅读：《通过 OCRmyPDF 实现在扫描版 PDF 中检索文字》

技巧五：批量制作 PDF

PDF 最「本职」的工作，估计就是作为一种万能的打印格式。不少人都有给文印店 PDF 文档的好习惯，即便店里设备再古董，打印出来格式也不太会出错。

不过对于制作 PDF 的人来说，可能就要花点力气了，遇上期末、项目截止时文档较多，还挺麻烦。**下面的两个自动化方法可以一次性把各类常见文档转换成 PDF，方便打印。**



各种格式的文章转换成 PDF

文本文档转换成 PDF

不含图片的文档主要有 3 类[5]

：

1. Docx 文档
2. HTML 文件
3. RTF 富文本文件

它们转换起来比较容易，在 macOS 上不用下载任何第三方工具，全靠系统工具就能解决。动作我已经做好，大家可以

直接下载，解压后双击安装到右键菜单就能用了。

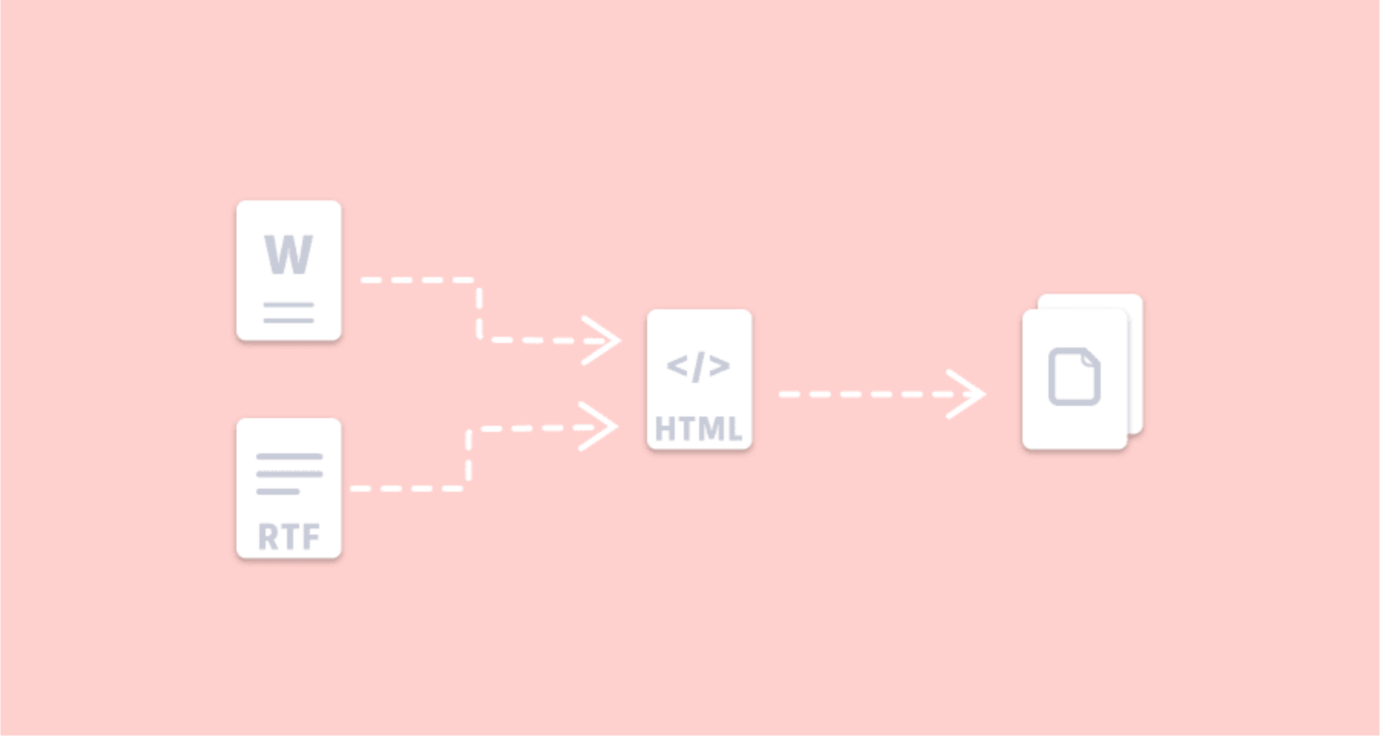
> 动作下载 🔗

注：本动作对于 RTF 和 HTML 文件支持良好，部分在 Docx 文档可能丢失格式。

稍微留意上面的视频，你会发现，**我这个动作不仅可以批量转换 PDF，而且 3 类文档通吃，一起选中也照样处理掉**。其实背后的原理挺简单的，就是根据不同文件格式、套用不同的系统命令[6]

。具体的代码我们不展开，只需简单了解一下上面 3 类文件的转换过程，以后遇到需要转格式的情况，你也可以对症下药：

1. Docx 文档：先用 `textutil` 命令生成 HTML，再用 `cupsfilter` 命令转换成 PDF。
2. RTF 文件：先用 `textutil` 命令生成 HTML，再用 `cupsfilter` 命令转换成 PDF。
3. HTML 文件：直接用 `cupsfilter` 命令转换成 PDF。



3 类文档的转换流程

含图片的文档转换成 PDF

如果是嵌有图片的文档，上面的方法就不起作用了，不过我们仍然有一个原生方案。

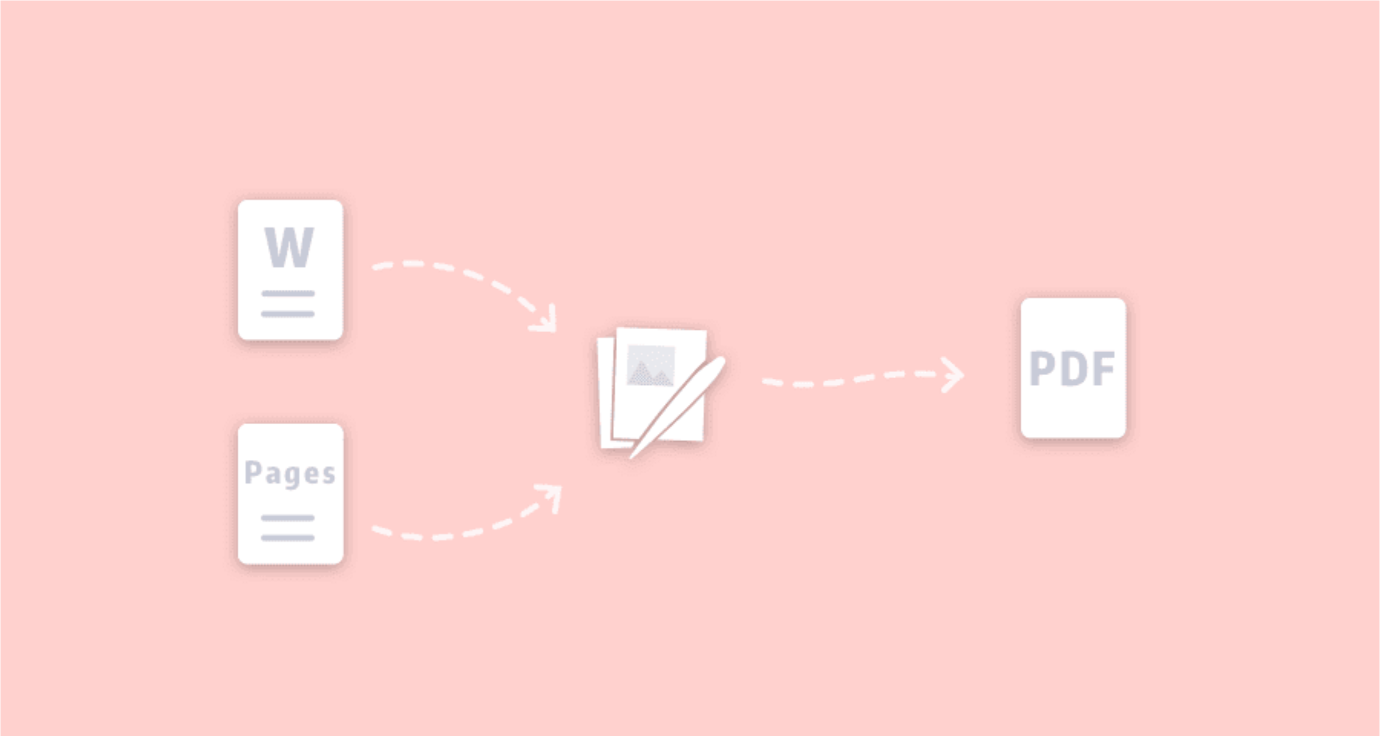
macOS 的老玩家可能还记得，可以安装 Office 套件，然后搭配 Automator 实现 Word 到 PDF 的转换。**其实有个很取巧的办法，完全用不着 Office，用免费的 Pages 就可以生成 PDF，而且通吃 Docx 和 Pages 文档。**

和前几节一样，这次动作还是 Automator Services，装好后同样会出现在右键菜单里。**一次性选中几个 Docx 和 Pages 文档，然后运行这个动作，稍等一会儿几份 PDF 就导出来了。**

> 动作下载 🔗

就像视频所演示的，只要你愿意，Docx 和 Pages 文档混着一起转换也行，照样能够打印成功。

动作原理并不复杂，只是把日常的「导出」操作写进了 AppleScript 脚本里面；有意思的地方在于，Pages 是可以打开预览（但是一些 Word 专有样式会丢失）Docx 文档的，所以不需要进行任何优化就可以让 AppleScript 像处理原生 Pages 文档一样打印 Docx——反正就是所见即所得嘛，没有什么深奥的东西。



通过 Pages 打印带图片的文档

小结

PDF 固然是一种方便的预览格式，而「看」以外的操作，往往不是它的特长。条件允许的话，应该选择 Word、Markdown 等更适宜的文档格式。

由于种种原因，我们仍然需要用到 PDF，而且涉及从页面调整到文本搜索等不属于其本职的操作。遇到这些情况，就需要一些巧劲儿了。